Dawei Li

Email: dal034@ucsd.edu Phone:858-291-9957 LinkedIn: Dawei Li | LinkedIn Website: Dawei Li | David-Li0406.github.io

EDUCATION

University of California, San Diego (UCSD) – San Diego, US **09/2022 - 03/2024 (expected)**

- Master of Science in Data Science, Overall GPA: 3.89/4.0, Major GPA: 3.89/4.0

Beijing Language and Culture University (BLCU) - Beijing, China **09/2018 - 06/2022**

- Bachelor of Engineering in Computer Science, Overall GPA: 86.4/100, Major GPA: 87.0/100
- Awards:

Outstanding Graduate Thesis Award (top 5%) at BLCU **06/2022**
The Meritorious Prize winner in "MCM (Mathematical Contest in Modeling)" **02/2021**
The 1st Prize winner (top 8%) in "Beijing Mathematical Contest in Modeling" **09/2020**
The 1st Prize winner (provincial level, top 10%) in "Contemporary Undergraduate MCM" **05/2020**

SELECTED PUBLICATION

Multi-level Contrastive Learning for Scripts-based Character Understanding (EMNLP 2023) **10/2023**
Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning (BEA-ACL 2023) **05/2023**
Fine-grained Contrastive Learning for Definition Generation (AACL 2022 Oral) **09/2022**
C^3KG: A Chinese Commonsense Conversation Knowledge Graph (ACL 2022) **02/2022**

RESEARCH EXPIRENCE

Contextualization Distillation from Large Language Model for Knowledge Graph Completion **06/2023-10/2023**

- Analyzed the **static and succinct** characteristics of the existing textual data used in knowledge graph completion; proposed a **contextualization distillation**, a **plug-in-and play** framework to distill informative and high-quality context from the large language model to train smaller knowledge graph completion model

DAIL: Data Augmentation for In-Context Learning via Self-Paraphrase **03/2023-06/2023**

- Analyzed the main limitation of the current in-context learning methods in low-resource and low-available scenarios and propose a **self-paraphrase** mechanism which utilizes the individual paraphrase of each sample to do ensembling
- Extensive empirical evaluation shows that DAIL outperforms the standard ICL method and other ensemble-based methods in the **low-resource scenario**
- Explore the use of **voting consistency** as a **confidence score** of the model when the logits of predictions are inaccessible and get promising results

READ: Improving Relation Extraction from an Adversarial Perspective **02/2022-06/2023**

- Proposed a novel adversarial attack method to explore the relation extraction models' **learning preference** when both entity and context are given; Reaveled the **over-dependency** and **non-generalization** of these models toward entities
- Designed an **entity-aware virtual adversarial training** method to address the aforementioned issue; Adopted **separate accumulated vocabulary** to foster perturbation searching and **clean token leaving** to encourage RE models to leverage indirect relational patterns in context
- Extensive experiments show that compared to various adversarial training methods, our method significantly improves **both the accuracy and robustness** of the model
- Additionally, experiments on different data availability settings highlight the effectiveness of entity-aware virtual adversarial training in **low-resource scenarios**

Multi-level Contrastive Learning for Script-based Character Understanding **09/2022-04/2023**

- Analyzed the difficulties of character-centric understanding in narrative scripts from the perspective of **text length** and **text type**; proposed a unified **multi-view contrastive learning framework** for script-based character understanding
- Introduce the summary of scripts as a simple-style counterpart of the dialogue and conduct **dialogue-summary contrastive learning** to help models understand the fine-grained information in dialogue; Designed **in-batch contrastive learning** to prompt models to learn the long-term dependency feature of each character globally
- Conducted experiments on a series of character-centric understanding tasks to validate the effectiveness of our method; analyzed the results and found our proposed method not only brings improvement to the pre-trained baselines but is also **compatible** with the previous SOTA methods

Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning **10/2022-01/2023**

- Benchmarked a new task called **trans-lingual definition generation** that generates definitions of the target words in the learners' native language, to assist them to capture the meaning of the unfamiliar word better
- Utilized the trans-lingual ability of neural machine translation models to do trans-lingual definition generation in an **unsupervised way**; proposed the **contrastive prompt** method to solve the two types of error found in baselines

Dawei Li

Email: dal034@ucsd.edu Phone:858-291-9957 LinkedIn: Dawei Li | LinkedIn Website: Dawei Li | David-Li0406.github.io

- Conducted experiments in both **rich- and low-resource** settings; reported experimental results together with some hints we found of trans-lingual definition generation task to foster future research

Definition Generation with Fine-grained Contrastive Learning **03/2022-07/2022**

- Noticed the **under-specific problem** in the definitions generated by the current SOTA pre-trained encoder-decoder models and analyzed the reasons for that phenomenon
- Proposed a novel **fine-grained contrastive learning** method to align the representation of the word and definition, to prompt the model to capture the full semantic components of the target word; Applied our method in a T5 backbone
- Conducted experiments on three mainstream benchmarks and prove that the proposed method could generate more **specific and high-quality definitions** compared with several state-of-the-art models.

C^3KG: A Chinese Commonsense Conversation Knowledge Graph **08/2021-03/2022**

- Collected a large Chinese conversation dataset CConv with high-quality utterances and fine-grained annotations; built a novel **commonsense conversation knowledge graph** based on CConv and the ATOMIC knowledge graph
- Mapped the content of each sub-utterance to the head node of the knowledge graph and modeled the dialogue flows as the edges of the graph; created a commonsense conversation knowledge graph with entities at **different granularity levels** and **rich dialogue transfer relationships**
- Conducted overall evaluations of our knowledge graph, including node evaluation and edge evaluation; Improved the performance of the baseline models with the knowledge triplets sampled from our knowledge graph

INTERNSHIP EXPIRENCE

Machine Learning Engineering Intern – AI Lab of Xiaomi, Beijing **08/2021 - 06/2022**

Knowledge Graph Annotation and Visualization Platform

- Built a search engine based on **Flask** and **Elasticsearch** to retrieve knowledge triplets for annotation; created APIs for node adding, deleting, modifying of the target knowledge graph
- Created data display page and data annotation tools components; implemented node hiding and hierarchical indexing functionality on the data display page to dynamically present annotation nodes and enhance page conciseness

Mental Support Conversational Recommendation Chatbot

- Sampled the triplets in C3KG to simulate users' **emotional cause chains**; use sampled data as the training data of the conversational recommendation system and built a **progressive mental support conversational recommendation chatbot** to reason user's emotional causes and provide suggestions
- Reproduced a production version code of the conversational recommendation model with TensorFlow; optimized and deployed the model in the cloud server; created an API to call the multi-turn model
- Built a demo to interact with the backend model; built a dialog component with **Vue.js**, including message boxes, input boxes, and interactive buttons; created a visualization box based on **D3.js** to interact with the model dynamically

Research Intern - Beijing Advanced Innovation Center for Language Resources- Beijing, China **06/2020 - 04/2021**

Methods for Definition Modeling of Multiword Expressions (MWEs)

- Developed a Scrapy-based multi-process crawler that utilized the producer-consumer model to manage the proxy IP pool; retrieved the number of Google search results of each phrase to calculate its **Multi-word Expression Distance**
- Integrated Multi-word Expression Distance knowledge into the BERT in the fine-tuning stage via **multi-task learning** to do definition generation of MWEs

PROJECTS

NLP Research Hotspots Analysis with Structural and Unstructured Data **09/2023 – 12/2023**

- Crawled information of NLP papers in recent three years from semantic scholar API; extracted keywords from the abstract of each paper and conducted **unsupervised topic clustering**; analyzed the authors' collaboration relationship with Neo4j; analyzed the combination of different topics with co-occurrence heatmap

Diff-Transferer: Any-Speaker Adaptive Text-to-Speech with Diffusion **01/2023 - 04/2023**

- Proposed Diff-Transferer, an any-speaker text-to-speech model with a **shallow diffusion mechanism**; stabilized the traning process of the diffusion model and reduced the number of training steps needed to reach convergence

Comparison of Transformer Models from Topological Perspective **04/2023 - 06/2023**

- Extract mention representations from the whole sentence encoding outputted by each transformer model (E.g. BERT, GPT2) and build **mapper graphs** as the model's topological summary; analyzed the mapper graph and determine the consistency of graph similarity between two models, corresponding to their **architectural similarity**