Zhengnan Xie

312-889-2420 | zhengnanx@email.arizona.edu | linkedin.com/in/zhengnan-xie-98a788117/ | github.com/zhengnanx

Education

University of Arizona — Tucson, AZ
Master of Science in Information(expected May 2021) GPA: 4.0 — *Aug. 2018 – May 2021*

Shanghai International Studies University — Shanghai, China
Bachelor of Arts in Linguistics GPA: 3.75 — *Aug. 2014 – June 2018*

Experience

Natural Language Processing Research Assistant — November 2018 – current
University of Arizona — *Tucson, AZ*

- **WorldTree Project:**
- Annotated a corpus of science exam questions of 7787 questions for fine-grained multi-class classification tasks
- Generated data entries (10,000 entries in total) for a semi-automatic science knowledge base corpus
- Generated structured explanations for around 2000 questions
- **Space Situational Awareness–Information Extraction task:**
- Authoring linguistic rules using Odinson language for a rule-based information extraction task
- Using python to postprocessing the extracted information
- Running BERT-NER model to compare performance between a rule-based system and a neural model
- **Question Classification tasks:**
- Generating data for fine-grained multi-class classification problems that scale to hundreds of classification labels

Web Development Intern — August 2019 – November 2020
Coeur d'Alene Online Language Resource Center — *Tucson, AZ*

- Using Sequalize and GraphQL to build/query/connect the database
- Using Hasura for auth backend and database backend
- Developing a full-stack web application using React, PostgreSQL, and Docker container for deployment

SemEval2018 Emoji Prediction — May 2019
University of Arizona

- Crawling twitter data for Emoji Prediction Tasks
- Built a predictive model for training and predicting
- Visualizing the model performance in a confusion matrix using sklearn library in python

Sentiment Analysis on 15 emotions — Spring 2019
University of Arizona

- Preprocessing tweets data by cleaning up the special chars
- Utilizing word2vec word embedding for the training
- Building a neural networks with bi-directional GRUs for the training

Publications

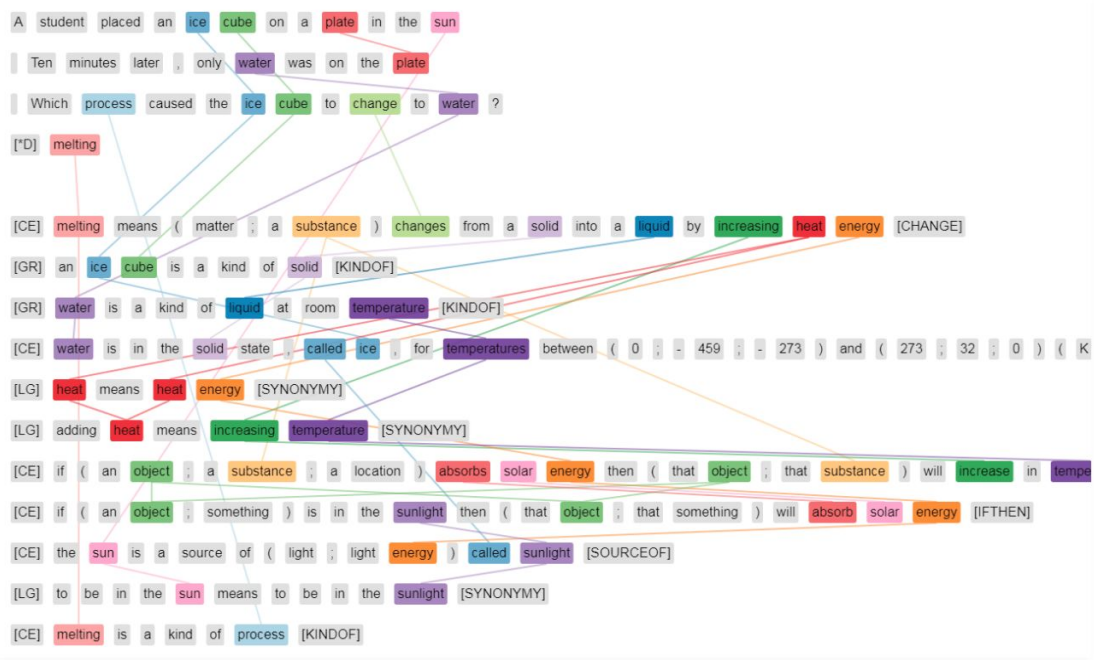
Multi-class Hierarchical Question Classification for Multiple Choice Science Exams | *LREC* — 2020
Dongfang Xu, Peter Jansen, Jaycie Martin, Zhengnan Xie, Vikas Yadav,Harish Tayyar Madabushi, Oyvind Tafjord and Peter Clark.

WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference | *LREC* — 2020
Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright,Steven Marmorstein, Peter Jansen.

Extracting Space Situational Awareness Events from News Text | *In Submission* — 2020
Zhengnan Xie, Peter A. Jansen, Moriba K. Jah

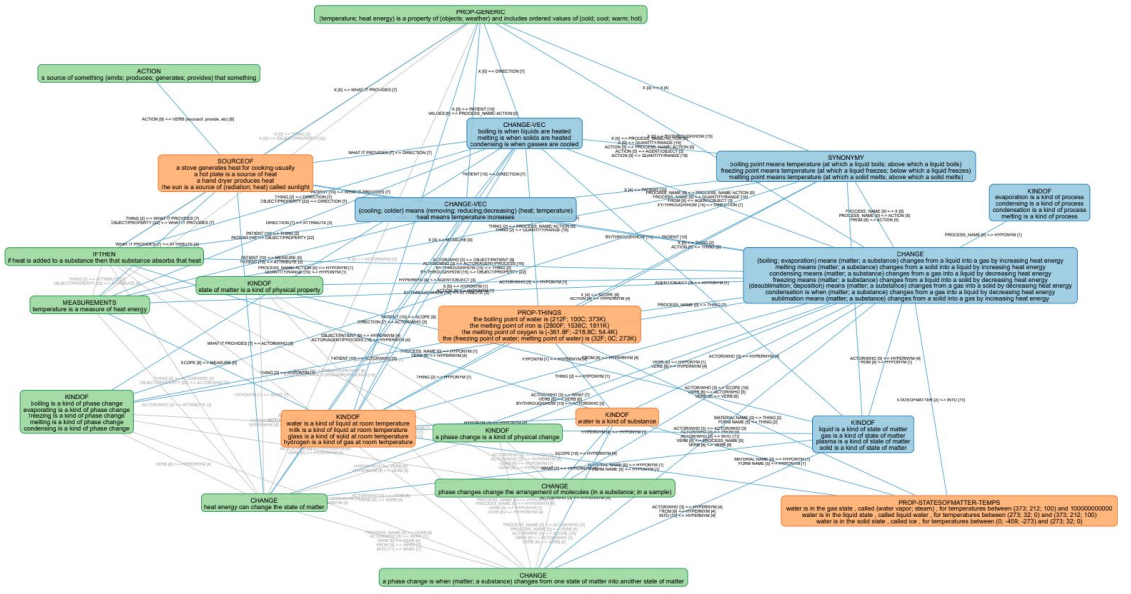
Technical Skills

Programming: Python, HTML/CSS, SQL, Java
Frameworks: React, Node.js, Odinson
Developer Tools: Git, Docker, VS Code, Linux, High Performance Computing, Hasura
Libraries: pandas, NumPy, Matplotlib, Sklearn, Keras
Languages: Chinese, English



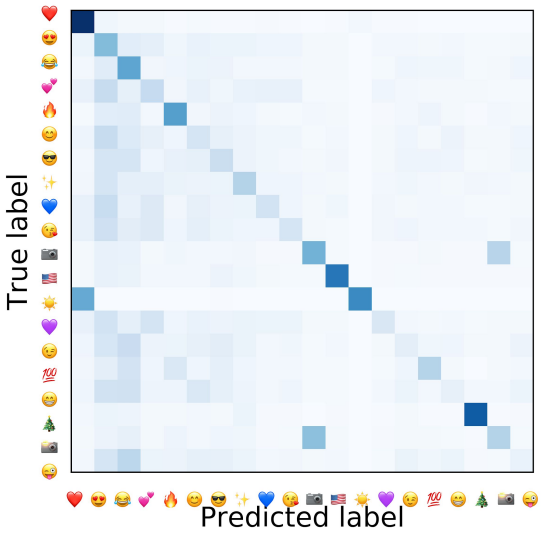
WorldTree V2.1

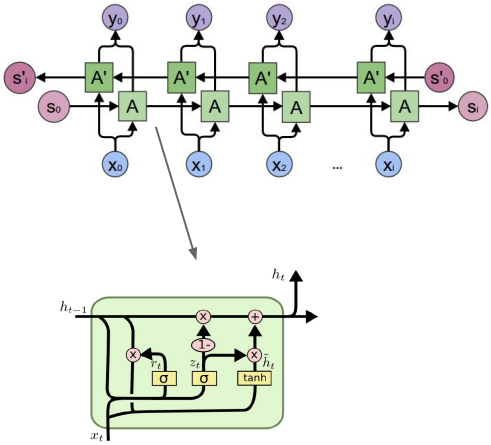
Assisted in generating a detailed semi-structured ontology of tables in the science domain, and applying this to generating detailed explanations (in the form of lexically-connected explanation graphs) for thousands of science exam questions



WorldTree V2.1

Experienced in abstracting complex patterns in graphs, and expressing these in simpler, interpretable rule-based formalisms



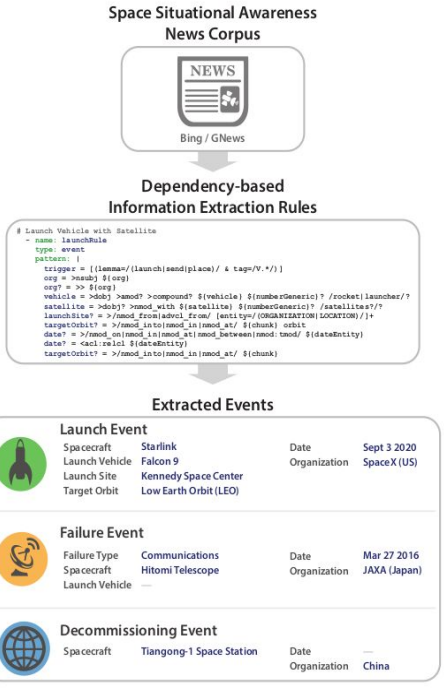


Emoji Prediction

Created a predictive model for determining emojis associated with each tweet from the SemEval2018 emoji prediction task. Overall F1 score was 33(35.99 scores the 1st), with low prediction error (as shown in the confusion matrix above)

Sentiment analysis

I built a bidirectional gated-recurrent-unit (GRU) neural network using word2vec to classify 15 classes of fine-grained sentiments(anger, anticipation, disgust, fear, joy, love, optimism, etc.) for social media data.

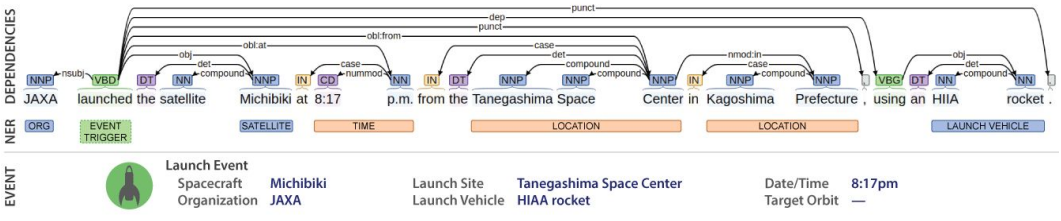


Space Event Information Extraction

I developed a series of 60 high-confidence (high precision/low-recall) and lower-confidence (lower-precision high-recall) information extraction rules for extracting important space events, such as launches and spacecraft failures, from news articles.

The extraction rules were expressed as a combination of syntactic dependencies and surface forms in the Odinson framework, and tailored to this low resource domain.

My overall extraction performance approached 80%, significantly exceeding a large language model (BERT) baseline by 10 points.



Astronomy / Celestial Events
- Planetary/Stellar Features
- Natural Cycles and Patterns
- Planetary/Stellar Distances
- Orbits

Earth Science
- Human Impacts on the Earth
- Weather
- Geology
- Outer Structure (Atmosphere/Hy
- Inner Structure (Crust/Mantle/C

Energy
- Properties of Light
- Converting Energy
- Electricity
- Sound Energy
- Potential/Kinetic Energy

Matter
- Chemistry
- Measurement
- Changes of State
- Properties of Materials
- Physical vs Chemical Ch
- Mixtures

Safety
- Safety Procedures
- Safety Equipment

Scientific Method
- Components of Inference
- Graphing Data
- Scientific Models

Other
- History of Science

Forces
- Gravity
- Friction
- Speed/Velocity
- Mechanical Energy
- Newton's Laws

Life Science
- Life Functions
  - Features and their Functions
    - Cellular Biology
    - Animal Features and Functions
    - Plant Features and Functions
      - Photosynthesis
      - Reproduction/Pollination
        - Seed Dispersal
      - Leaves
      - Roots
  - Environmental Effects on Development
  - Responses to Environment Changes
  - Basic Life Functions
- Interdependence/Food Chains
- Reproduction
- Adaptations and the Environment
- Continuity of Life/Life Cycle

Science Questions Classification

experienced in high-accuracy data generation pipelines for fine-grained multi-class classification problems that scale to hundreds of classification labels

Coeur d'Alene Language Resource Web Page Development

I joined the group to redevelop the website using Node.js, React, GraphQL, and mySQL(community edition) to allow increased functionality and more responsive interaction.

We use Docker to deploy our model on the server.