ANASTASIA KUZNETSOVA

(812) 558 80 55 | anakuzne@iu.edu
https://github.com/ana-kuznetsova/
https://www.linkedin.com/in/anastasia-kuznetsova-2bb66b116/

EDUCATION

Indiana University, Bloomington *Aug 2019 - May 2024*
PhD student in Computer Science and Computational Linguistics GPA: 3.72

Courses: Deep Learning for speech processing, Advanced Natural Language Processing, Reinforcement learning, Machine Learning

NRU Higher School of Economics *Sep 2017 - Jul 2019*
Master of Arts
Computational linguistics Moscow, Russia

TECHNICAL STRENGTHS

Programming skills: Python (TensorFlow, PyTorch, Flair, NLTK, Scikit learn, OpenAI gym), Bash (Kaldi, HFST), JavaScript, R, C++.
Versioning control: Git
Data Bases: SQL.
Container Tools: Docker.

CURRENT PROJECTS

Reinforcement Learning based Speech Enhancement
Ongoing project explores the possibilities of exploiting RL algorithms in speech enhancement area in low-data setting. Current phase involves implementation and testing of the existing RL-based speech denoising techniques from published papers.

A reinforcement-learning approach to curriculum generation for ASR
The project suggests an approach to mitigate the lack of training data by employing an Automated Curriculum Learning paradigm in combination with bandit approach inspired by the *reinforcement learning* domain. The implemented bandit framework shows a 10% WER and 27% CER improvement over the baseline model and the potential of the further investigation of RL algorithms in ASR.

Speaker Identification system
The system is implemented for the multi-speaker Spanish data set. The speakers are differentiated based on the distance between two speaker embeddings classified as similar or dissimilar. *Siamese* network architecture allowed not only to separate speaker embeddings but leverage *data augmentation* as well. My role in this project included implementation of the neural architectures in Keras and Tensorflow.

BPE weighting of morphological analyser for Paraguayan Guaranai
Weighting of the morphological analyser based on finite-state technology using Byte Pair encoding algorithm.

Google Summer of Code: Machine Translation for Guarani–Spanish language pair
Apertium is an Open Source machine translation platform focusing on under-resourced and marginalized languages. The project includes the construction of FST-based morphological analyser, Guarani-Spanish bilingual dictionary (bidix) and Transfer rules.
`http://wiki.apertium.org/wiki/User:Anakuznetsova/GSOC_2018_Guarani_Spanish`

Scalable Text Analyser Web Application The web application has 4 APIs communicating with each other through Celery messaging queues. Each API service runs a separate text analysing task: Term Extraction, Text Classification, Named Entity Recognition, Calculation of Readability metrics.

Named Entity Recognition for Russian Popular Science corpora
Rule-based entity extraction using Tomita parser. I leveraged corpora manually annotated with the names of the scientists and wrote the rules achieving 0.48 accuracy in rigid time constraints.

Morphological Disambiguation for Paraguayan Guaraní
Morphological disambiguation task is being completed by using rule-based technology (Constraint Grammar formaism).

WORK EXPERIENCE

Indiana University, Bloomington, USA Sep 2019–Present
Graduate Research Assistant

· Research in technologies for low-resourced languages with the focus on research in Automated Speech Recognition and Speech Enhancement.

MTS, Artificial Intelligence Department, Moscow, Russia Jan–Jul 2019
Software Developer

· Development of skills for a Smart Speaker as well as developing auxiliary NLP tools for the team of Computational Linguists: text classifier which detected hate speech, adult content, and the other tool for classification of similar user requests.

EXTRA-CIRRUCULAR

Google Summer of Code 2019, 2020
Mentor for Apertium

· Mentoring of machine translation projects for Medumba–French and Hindi–Punjabi language pairs.

Google Code-In 2018, 2019
Mentor for Apertium

· Mentoring for Apertium Open Source organization. Helped younger students in completing coding tasks.

Google Summer of Code 2018
Student Developer in Apertium

· Developed machine translation system for Guarani–Spanish language pair.

RESEARCH PAPERS

A finite state morphological analyser for Paraguayan Guaraní (LREC). Co-authored with F. M. Tyers (Under review).

A finite state morphological analyser for Evenki (LREC 2020). Co-authored with F. M. Tyers and A. Zueva. `https://www.aclweb.org/anthology/2020.lrec-1.314`

LANGUAGES

Russian (native), Portuguese (fluent), Spanish (fluent), Lithuanian (intermediate), Guaraní (basic knowledge)