ANASTASIA KUZNETSOVA

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

EDUCATION

Indiana University, Bloomington *Sep 2019 - May 2024*
PhD student in Computational linguistics
and Computer Science GPA: 3.825
Courses: Deep Learning for speech processing, Advanced Natural Language Processing, Reinforcement learning, Graduate Seminar in Computational Linguistics

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

TECHNICAL STRENGTHS

Programming skills: Python (TensorFlow, Keras, PyTorch, Flair, NLTK, Scikit learn), Bash (Kaldi, HFST), JavaScript, Django, R, C++.
Versioning control: GitHub, GitLab
Data Bases: SQL.
Container Tools: Docker, Singularity, Celery.

CURRENT PROJECTS

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`

Curriculum learning for low-resourced languages
The project is addressing the problems of Automated Speech Recognition for marginalized languages in *low data scenarios*, exploiting the idea of Curriculum learning to train ASR systems and applying it on Tatar (Turkic language of Russia) and Breton (regional language of France) with less than 20 hours of audio.

Speaker Identification system
Speaker Identification on accented speech data which leverages techniques of *data augmentation* and *speech enhancement*. The approach is based on the distance between speaker embeddings. My role in this project is coding neural network architectures in Tensorflow and Keras such as LSTMs and Siamese networks.

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.

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

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, speech processing areas.

MTS, Artificial Intelligence Department, Moscow, Russia Jan–Jul 2019
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 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.

PUBLICATIONS

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

(Accepted) A finite state morphological analyser for Evenki (LREC). Co-authored with F. M. Tyers and A. Zueva.

LANGUAGES

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