Anton Belyy

anton.belyy@gmail.com · avbelyy.github.io
Scholar · GitHub · LinkedIn

Education

Sep '19 — May '22 | **M.Sc.Eng. in Computer Science (NLP)**, Johns Hopkins University, USA
GPA: 4.0. Advisors: Benjamin Van Durme (Primary), Vladimir Braverman (Secondary)

Sep '13 — May '18 | **B.Sc. in Informatics and Applied Mathematics**, ITMO University, Russia
GPA: 3.6. Advisors: Andrey Filchenkov (Primary), Konstantin Vorontsov (Secondary)
Thesis: Construction and Quality Evaluation of Heterogeneous Hierarchical Topic Models
- **Distinguished thesis award** (given to 2 out of approx. 50 undergraduates)

Publications and Preprints

1. **Logical Satisfiability of Counterfactuals for Faithful Explanations in NLI.**

   Sia, S., Belyy, A., Almahairi, A., Khabsa, M., Zettlemoyer, L., & Mathias, L. (2022). Accepted to the *Beyond Bayes Workshop within ICML2022*. [preprint]

2. **Human Schema Curation via Causal Association Rule Mining.**

   Weber, N., Belyy, A., Holzenberger, N., Rudinger, R., & Van Durme, B. (2022). In *Proceedings of The 16th Linguisitic Annotation Workshop (LAW-XVI) within LREC2022*, pp. 139-150. [paper] [code] [demo] [data]

3. **Best K-best: Efficient Item Selection for Rapid Data Annotation.**

   Belyy, A., Huang, C.-Y., Andreas, J., Platanios, E. A., Thomson, S., Shin, R., Roy, S., Chen, C., & Van Durme, B. (2022). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 114-126. [paper] [poster] [slides] [talk]

4. **InFillmore: Frame-Guided Language Generation with Bidirectional Context.**

   Ou, J., Weir, N., Belyy, A., Yu, F., & Van Durme, B. (2021). In *Proceedings of the 10th Conference on Lexical and Computational Semantics*, pp. 129-142. [paper] [poster] [slides] [talk] [demo]

5. **Script Induction as Association Rule Mining.**

   Belyy, A., & Van Durme, B. (2020). In *Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events*, pp. 55-62. [paper] [slides] [talk] [code]

6. **Improved Evaluation Framework for Complex Plagiarism Detection.**

   Belyy, A., Dubova, M., & Nekrasov, D. (2018). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 157-162. [paper] [poster] [code]

7. **Framework for Russian Plagiarism Detection Using Sentence Embedding Similarity and Negative Sampling**.

   Belyy, A., & Dubova, M. (2018). In *Proceedings of the 24th International Conference on Computational Linguistics and Intellectual Technologies*, Issue 17, pp. 96-109. [paper] [slides] [code]

8. **Quality Evaluation and Improvement for Hierarchical Topic Modeling**.

   Belyy, A., Seleznova, M., Sholokhov, A., & Vorontsov, K. (2018). In *Proceedings of the 24th International Conference on Computational Linguistics and Intellectual Technologies*, Issue 17, pp. 110-123. [paper] [slides]

Work Experience

Jun '22 – Present | **Machine Learning Engineer** in **Knowledge Platform**, Apple, USA
- Building ML systems for large-scale knowledge extraction from unstructured data feeds

Sep '17 — Aug '19 | **Senior Data Scientist** in **Compliance Risks** and **AI lab**, Tochka Bank, Russia
- **Risk scoring**: vectorized new data sources for 200K+ bank clients and 50M+ transactions, generated temporal/spatial features using Hadoop/Spark to improve scoring accuracy by 10%
- **Communication analysis**: built intent recognition models to classify 90% customer inquiries
- **Call center planning**: using OR-Tools, automated CC planning and improved accuracy by 10%
- **ML culture**: interviewed and mentored 3 junior ML engineers, designed internal ML guidelines

Mar '17 — Aug '17 | **Machine Learning Engineer**, Antirutina, Russia
- **Tender anomaly detection**: developed clustering algorithms to identify bidding anomalies, allowing to discover bid-rigging behavior on auctions with contract amount exceeding $3B
- **Precise IE**: designed information extraction pipelines for precise identification of vendor codes, volumes and quantities of goods from unstructured and diverse vendors' price lists

Oct '15 — Oct '16 | **Software Engineer** in **Ads**, VK.com, Russia
- **URL fraud**: built service to periodically detect malicious URL redirect changes in VK ads
- **Click fraud**: built ML models to detect users that generate fraudulent clicks in VK ad network. Model was deployed semi-automatically and helped recover up to 3% monthly ad revenue
- **Ads search**: launched moderator search interface (incl. full-text search) over 30M+ VK ads
- **Ads scoring**: implemented advertiser ranking for faster moderation of top-10% clients

Research Experience

Sep '19 – May '22 | **Graduate Research Assistant**, Johns Hopkins University, USA
- **Semantic data mining**: proposed novel ARM-based algorithm for script induction [paper], built SchemaBlocks, Scratch-like annotation interface for complex event scenarios [paper]
- **Knowledge graph completion**: building a human-in-the-loop KG completion system using entity linking, rule learning and data mining over million-scale knowledge graphs
- **Text generation**: built demo for InFillmore, our FrameNet frame-guided NLG model [paper]

Jun '21 — Aug '21 | **Research Intern** in **Semantic Machines**, Microsoft Research, USA
- Built guided annotation interface to help label semantic parsing data 35% faster [paper]

Mar '18 – Jun '18 | **Research Intern**, Université Grenoble Alpes, France
- Built extreme multi-class classification systems using Pegasos and MIPS algorithms [report]

Mar '17 — Mar '18 | **Undergraduate Research Assistant**, ITMO University, Russia

Contributions to the areas of **plagiarism detection** and **exploratory search**:
- Novel evaluation metric for external plagiarism detection [paper]
- Framework for external plagiarism detection in Russian [paper]
- Hierarchical topic modeling for exploratory search over heterogeneous sources [paper]
- Topic-model driven exploratory search engine system [code] [demo]

Teaching Experience

Jan '21 — May '21 | **Introduction to Algorithms 601.433/633 (Head TA)**, JHU (100+ students)
- Managed 9 CAs and 1 TA, created homework and exam problems, held weekly office hours

Sep '17 — Aug '19 | **Natural Language Processing (TA)**, Coursera (40,000+ students by Sep '19)
- Answered 200+ students' questions, helped create homework and project assignments

Languages and Technologies

Languages	Python (proficient); JavaScript, bash (intermediate); C#, C++, Haskell, x86 assembly (coursework)
Technologies	pandas, sklearn, XGBoost, pytorch, faiss, nmslib; Docker, *SQL, MongoDB, Lucene, Hadoop/Spark