Vilém Zouhar

vilem.zouhar@gmail.com | vilda.net | github.com/zouharvi

EDUCATION

ETH Zürich 2022-present
PhD in Computer Science in Mrinmaya Sachan's lab

Saarland University + Groningen University (2 years) 2020 - 2022
MSc. in Language Science and Technology (funded scholarship)
Double degree programme
Thesis Shrinking Knowledge Base Size: Dimension Reduction, Splitting & Filtering

Charles University in Prague (3 years) 2017 - 2020
BSc. in Computer Science (+ graduate-level classes)
Specialization in Computational linguistics
Thesis Enabling Outbound Machine Translation

TECHNICAL KNOWLEDGE

Programming	Python, JS/TS, Rust, C/C++, R
Toolkits	PyTorch, Scikit, Numpy, Huggingface, Marian NMT, Fairseq, Matplotlib
Misc.	Linux (long-term user, GPU cluster etc), visualization, typesetting (pandoc, LaTeX)

LANGUAGE PROFICIENCY

Czech	Native
English	C2 *(iBT TOEFL 118/120)*
German	B2 *(in development)*
Others	bits of random languages (linguistic curiosity)

TEACHING

Large Language Models (materials) summer semester 2023

Neural Networks Implementation and Application class (materials) winter semester 2021

Statistical Natural Language Processing class (materials) summer semester 2021

WORK EXPERIENCE

Spoken Language Systems group (student research assistant) (14 months) 2021-2022
- University of Saarland (Germany)
- Information retrieval efficiency through dimensionality reduction (code)
- Language modelling with an external source of information

Institute of Formal and Applied Linguistics (student research assistant) (3 years) 2019-2022
- Charles Univeristy (Czech Republic)
- Machine translation related projects
- Bergamot project (in-browser MT)
- Psycholinguistic project consultation
- Miscellaneous research tasks

Previo (intern software dev, multilayer CMS) (3 months) summer 2018

BIM Project (intern software dev, C++/Boost and C#) (3 months) summer 2017

Web development (PHP/JS/TS/HTML/CSS stack) 2015-2017

SELECTED PUBLICATIONS

Google Scholar Semantic Scholar

Poor Man's Quality Estimation:
Predicting Reference-Based MT Metrics Without the Reference
EACL 2023
Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim,
Tom Kocmi, Yuchen Eleanor Jiang, Mrinmaya Sachan

Sentence Ambiguity, Grammaticality and Complexity Probes
BlackboxNLP 2022
Sunit Bhattacharya,[=] Vilém Zouhar,[=] Ondřej Bojar

Stroop Effect in Multi-Modal Sight Translation
Preprint
Sunit Bhattacharya, Vilém Zouhar, Věra Kloudová, Ondřej Bojar

Fusing Sentence Embeddings Into LSTM-based Autoregressive Language Models
Preprint
Vilém Zouhar, Marius Mosbach, Dietrich Klakow

Knowledge Base Index Compression via Dimensionality and Precision Reduction
ACL Spa-NLP 2022
Vilém Zouhar, Marius Mosbach, Miaoran Zhang, Dietrich Klakow

EMMT: A simultaneous eye-tracking, 4-electrode EEG and audio corpus for multi-modal reading and translation scenarios
Preprint
Sunit Bhattacharya, Věra Kloudová, Vilém Zouhar, Ondřej Bojar

Neural Machine Translation Quality and Post-Editing Performance
EMNLP 2021
Vilém Zouhar, Ondřej Bojar, Martin Popel, Aleš Tamchyna

Providing Backtranslation Improves Users Confidence in MT, Not Quality
NAACL 2021
Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill,
Frédéric Blain, Marina Fomicheva, Lucia Specia, Lisa Yankovskaya

Artefact Retrieval: Overview of NLP Models with Knowledge Base Access
AKBC CSKB 2021
Vilém Zouhar, Marius Mosbach, Debanjali Biswas, Dietrich Klakow

Sampling and Filtering of Neural Machine Translation Distillation Data
NAACL SRW 2021
Vilém Zouhar

Leveraging Neural Machine Translation for Word Alignment
PBML 116
Vilém Zouhar, Daria Pylypenko

WMT20 Document-Level Markable Error Exploration
WMT20
Vilém Zouhar, Tereza Vojtěchová, Ondřej Bojar

Extending Ptakopět for MT User Interaction Experiments
PBML 115
Vilém Zouhar, Michal Novák

Outbound Translation User Interface Ptakopet: A Pilot Study
LREC 2020
Vilém Zouhar, Ondřej Bojar

A Collection of Machine Learning Excercises
teaching materials
50 pages of ML tasks in R; full version available per request
Awarded Student Faculty Grant at Charles University

SERVICE

Reviewing: EACL 2023, SVRHM 2022, CoNLL 2022, AACL-IJCNLP 2022

CSRR: Workshop on Commonsense Representation and Reasoning (organizer) ACL 2022

Evaluation committee member for granting university accreditations (4 times)
in the Czech Republic 2020-present

Institute of Formal and Applied Linguistics, Charles University 2019-2022

- NER Presentation at NKÚ (supreme audit office)
- Department presentation at Open Days
- ELITR project coordination at a hackathon UniHack

ACADEMIC MISC.

- ryanize-bib: a tool to improve BibTeX standards 2023
- MachineTranslate.org: open guide to machine translation (contributor) 2021-present
- Stolen Subwords: Importance of Vocabularies for Machine Translation Model Stealing 2022
- Poetry, Songs, Literature, Legalese and Translationese: Automated Sentence Complexity Perspective 2022
- Generator of paper titles based on scientific abstracts 2022
- Pandemic Crisis Communication: Automatic Classification of Interviews With Experts 2021
- Fact Learning with Adaptive Color Palette: Effect of Stimuli-Independent Hints 2021
- Hyperparameters of RNN Architectures: for POS Tagging using Surface-Level BERT Embeddings 2021
- Deep Molecule: Quantitative Structure-Property Relationships 2021
- SlowAlign: IBM model-based word aligned with extra features and heuristics 2020
- SlowAlign Displayer: Quick online word-alignment visualization tool 2020
- MosQEto: Machine translation quality estimation data synthesis 2019
- Other small projects either for convenience, out of professional interest, or as a hobby, hosted at GitHub

EXTRA-CURRICULAR

Academic senate 2018-2020

- Member of the Academic Senate at Charles University Faculty of Mathematics and Physics
- Participation in Faculty meetings, communicating with students, introductory summer camp

Game jams 2015-present

- Several games programmed and presented in limited time, mostly Ludum Dare

Kasiopea 2017-2019

- Organization of Kasiopea, an annual coding competition for talented high school students

Music 2021-present

- Random Strum Pattern Generator