Vilém Zouhar

vilem.zouhar@gmail.com · vilda.net

🏛 Education

ETH Zürich   2022-present
PhD in Computer Science in NLPED lab
Advisors: Mrinmaya Sachan, Menna El-Assady
Thesis: Robustness-Aware Human-Centered NLP

Saarland University + Groningen University   (2 years) 2020 - 2022
MSc. in Language Science and Technology (scholarship)
Advisors: Dietrich Klakow, Gosse Bouma
Thesis Shrinking Knowledge Base Size: Dimension Reduction, Splitting & Filtering

Charles University in Prague   (3 years) 2017 - 2020
BSc. in Computer Science (+ classes in computational linguistics)
Advisor: Ondřej Bojar
Thesis Enabling Outbound Machine Translation

🔧 Technical Knowledge   **GitHub**

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

💬 Languages

▸ Czech: C2
▸ English: C2
▸ German: B2
▸ + linguistic curiosity

🧑‍🏫 Teaching

▸ Semestral project supervision   2023-present
▸ Large Language Models materials contribution   summer semester 2023
▸ Neural Networks Implementation and Applications tutoring & materials   winter semester 2021
▸ Statistical Natural Language Processing tutoring & materials   summer semester 2021
▸ A Collection of Machine Learning Excercises in R materials only

👷 Work Experience

▸ Amazon Machine Translation (research intern, quality estimation project)   (4 months) 2023

▸ Spoken Language Systems group (student research assistant)   (1 year) 2021-2022
  • University of Saarland, Germany
  • Information retrieval and language modelling efficiency

▸ Institute of Formal and Applied Linguistics (student research assistant)   (3 years) 2019-2022
  • Charles Univeristy, Czech Republic
  • MT- and psycholinguistics-related projects (Bergamot/in-browser MT, ELITR)

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

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

🐕 Service

▸ WMT Terminology Shared Task (organizer)   **2023**
▸ Reviewer: {EMNLP, AACL-IJCNLP, ACL, EACL} 2023, {SVRHM, CoNLL, AACL-IJCNLP} 2022
▸ CSRR: Workshop on Commonsense Representation and Reasoning (organizer)   ACL 2022
▸ Evaluation committee member for granting university accreditations in Czechia   (5 times) 2020-present

📑 Selected Publications

▸ **Tokenization and the Noiseless Channel** **ACL 2023**
Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Mrinmaya Sachan, Ryan Cotterell

▸ **A Formal Perspective on Byte-Pair Encoding** **ACL 2023**
Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, Ryan Cotterell

▸ **Re-visiting Automated Topic Model Evaluation with Large Language Models** **In review**
Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, Elliott Ash

▸ **Enhancing Textbooks with Visuals from the Web for Improved Learning** **In review**
Janvijay Singh, Vilém Zouhar, Mrinmaya Sachan

▸ **PWESuite: Phonetic Word Embeddings and Tasks They Facilitate** **In review**
Vilém Zouhar, Kalvin Chang, Chenxuan Cui, Nathaniel Carlson, Nathaniel Robinson, Mrinmaya Sachan, David Mortensen

▸ **Multimodal Shannon Game with Images** **Preprint**
Vilém Zouhar,⁎ Sunit Bhattacharya,⁎ Ondřej Bojar

▸ **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: An eye-tracking, 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**
V. Zouhar, M. Novák, M. Žilinec, O. Bojar, M. Obregón, R. L. Hill, F. Blain, M. Fomicheva, L. Specia, L. 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

▸ **Evaluating Optimal Reference Translations** **In review**
Vilém Zouhar, Věra Kloudová, Martin Popel, Ondřej Bojar

🧱 Misc. Projects

▶ Metaphor Preservation in Machine Translation and Paraphrasing **2023**
Report/essay on preserving metaphors in modern NLP

▶ Dorfromantik solver **2023**
A tool to optimize a game

▶ Ryanize bib **2023**
Tool to check for common bib best practice violations

▶ Poetry, Songs, Literature, Legalese and Translationese **2023**
Essay on Automated Sentence Complexity Perspective

▶ Stolen Subwords **2023**
Report on Importance of Vocabularies for Machine Translation Model Stealing

▶ Bilingual scientific abstracts corpus **2022**
Bilingual Czech and English abstracts of ÚFAL papers

▶ Machine Translate **2022**
Open resources and community for machine translation (contributor)

▶ Random Strum Pattern Generator **2022**
Generate random strum patterns to learn guitar strumming

▶ Deep Molecule QSPR **2021**
Collaboration with Nikola Kalábová on predicting key temperature points (e.g. boiling point) of molecules.

▶ Fact Learning with Adaptive Color Palette: Effect of Stimuli-Independent Hints **2021**
An experiment with enhancing fact learning experience in collaboration with Leander van Boven, Tianyi Li and Anjali Nair.

▶ Hyperparameters of RNN Architectures for POS Tagging using Surface-Level BERT Embeddings **2020**
Final project for neural network class.

▶ SlowAlign **2020**
IBM model-based word aligned with extra features and heuristics. Written in Rust.

▶ Slow Align Displayer **2020**
Creates quick graphs given word alignment (in the Pharaoh format).

▶ MosQEto **2019**
Collaboration with Ondřej Měkota on synthetising machine translation quality estimation data.

▶ Call for Menza **2019**
Aggregator of daily menus around the Faculty of Mathematics and Physics. Unmaintained since I moved out of Prague.

▶ SMAKE **2019**
Simple Markable And Keyword Extraction in Rust. Contributions by Petr Houška.

▶ TNTranslator **2019**
Translation Inspector for n-best list navigator task at Bergamot.

▶ ZimaDB **2018**
SQLite-like database implementation from scratch in C++ in collaboration with Petr Chmel.

▶ ASM Hell **2017**
Learn basic assembly instructions through a game (LD41 submission).

▶ Prolog KNN **2017**
Implementation of kNN in Prolog, a language the least suited for this.

📜 Extra-Curricular

▶ Academic senate at Charles University Faculty of Mathematics and Physics — 2018-2020
▶ Several **games** programmed and presented in limited time, mostly Ludum Dare — 2015-present
▶ Organization of **Kasiopea**, an annual coding competition for talented high school students — 2017-2019
▶ Electric guitar 🎸 — 2021-present