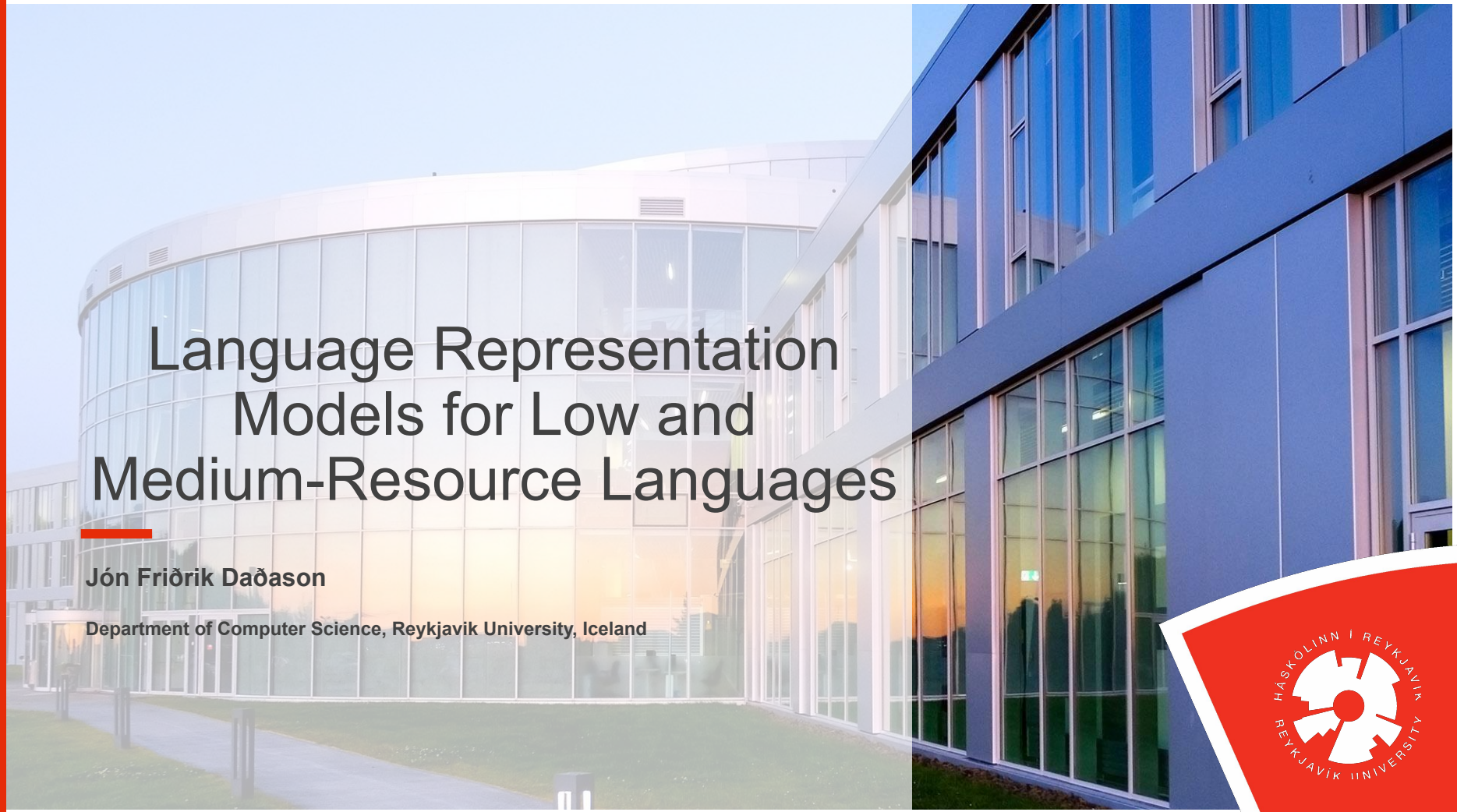

Language Representation Models for Low and Medium-Resource Languages

Jón Friðrik Daðason

Department of Computer Science, Reykjavik University, Iceland

Overview

- **Question**: How can we efficiently pre-train language models when language and computational resources are scarce?

- My research is focused on:
  - Subword tokenization algorithms (e.g., comparing BPE vs. Unigram)
  - Multilingual pre-training corpora (few vs. many languages)
  - Text filtering (rule-based and classifier-based)
  - Data-efficient pre-training tasks (e.g., MLM vs. RTD)
  - Data augmentation (e.g., back/machine translated text)

Results so far

- Subword tokenization
  - The choice of algorithm doesn’t appear to have a significant effect
  - Increasing the vocabulary size can improve downstream performance, but at a cost
- Multilingual corpora
  - Adding other Nordic languages to an Icelandic pre-training corpus improves performance for some tasks, but on average the performance is about the same
- Text filtering
  - Most commonly applied rules only filter out a handful of documents
  - Perplexity-based classifiers are highly successful

Thank you

- If you have any questions or comments, or are interested in collaborating with me, feel free to contact me!
- E-mail: jond19@ru.is