LuxemBERT: Simple Data Augmentation in Language Model Pre-Training for Luxembourgish
C. Lothritz, B. Lebichot, K. Allix, L. Veiber, T. F. Bissyandé, J. Klein, SnT, University of Luxembourg
C. Lefebvre, A. Boystov, A. Goujon, BGL BNP Paribas

Summary
• We train a BERT model for the Luxembourgish language named LuxemBERT
• We create several datasets for various NLP tasks in Luxembourgish to evaluate LuxemBERT: POS-tagging, NER, Intent Classification, News Classification, and Winograd Natural Language Inference
• LuxemBERT manages to beat mBERT in 6 out of 6 tasks as cased version, and in 5 out of 6 tasks as uncased version

Pre-Training
• We collect 6.1 million sentences from sources such as Wikipedia, RTL news station, a Luxembourgish chatroom, etc.
• We systematically translate non-ambiguous function words from a German dataset to Luxembourgish as means of data augmentation

Fine-Tuning
• To evaluate LuxemBERT’s performance, we train on 6 downstream tasks:
  • Part-of-Speech tagging
  • Intent Classification
  • Non-Trivial Intent Classification
• We compare LuxemBERT to mBERT as both cased and uncased versions

Results
• LuxemBERT outperforms mBERT in almost all tasks, often significantly
• Uncased models typically perform better than cased

Ablation Study
• We compare LuxemBERT to two more BERT models:
  • Lb BERT, trained on Luxembourgish sentences only
  • De/Lb BERT, trained on Luxembourgish and non-translated German sentences
• There is no clear winner for this dataset size

De/Lb BERT and LuxemBERT both usually outperform Lb BERT when less data is available
• There is, however, no obvious winner between De/Lb BERT and LuxemBERT