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

TruX - Trustworthy Software

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

Meaning (for readers) (English)	*There are 26 known isotopes, only two of which appear in nature.*
Sample text in auxilliary language (**German**)	Bekannt sind 26 Isotope, wovon nur zwei natürlich vorkommen.
Translated text for data augmentation (**pseudo-Luxembourgish**)	Bekannt sinn 26 Isotope, wouvun nëmmen zwee natierlech vorkommen
Ground-truth translation (Luxembourgish)	*Bekannt sinn 26 Isotopen, wouvun der nëmmen zwee natierlech virkommen*

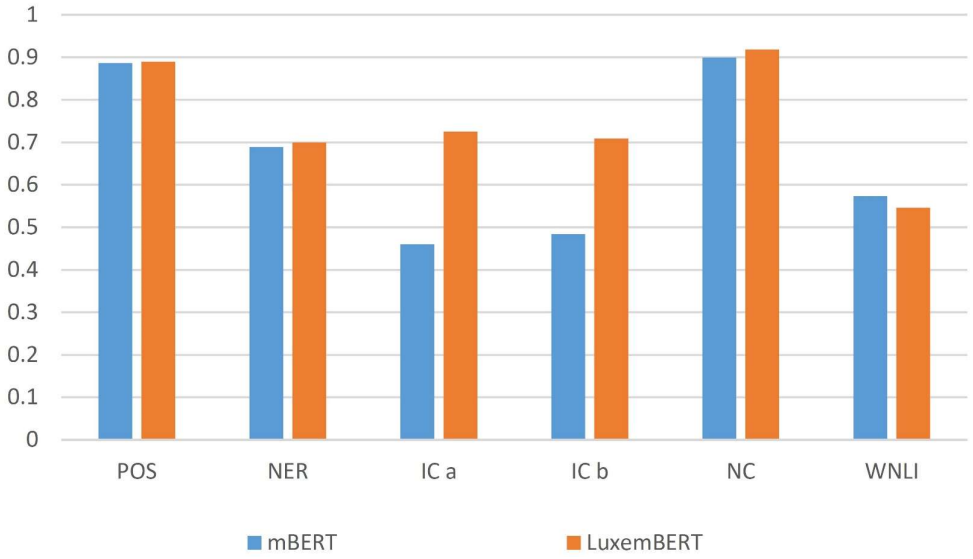
- We use the resulting 12.2 million sentences to pre-train our BERT model on MLM task

Fine-Tuning

- To evaluate **LuxemBERT**'s performance, we train on **6 downstream tasks**:
  - Part-of-Speech tagging
  - Intent Classification
  - Non-Trivial Intent Classification
  - Named Entity Recognition
  - News Classification
  - Winograd Natural Language Inference
- We **compare LuxemBERT to mBERT** as both cased and uncased versions

Results

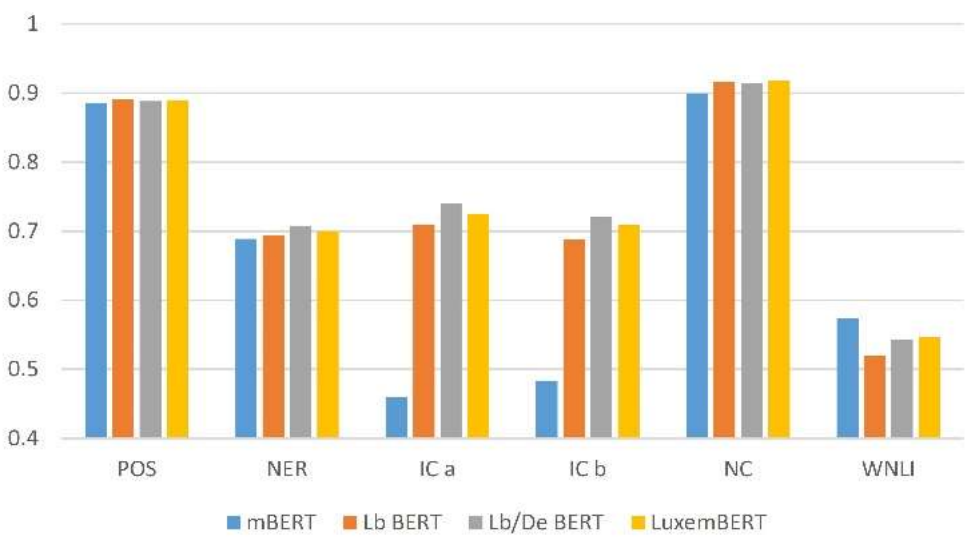
- **LuxemBERT outperforms mBERT** in almost all tasks, often significantly
- Uncased models typically perform better than cased



Performances of uncased mBERT vs uncased LuxemBERT tested on 6 downstream tasks

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



- We also train models **on smaller datasets (500k and 2.1M)** to evaluate our approach on lesser-resourced languages
- **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